American Journal of Epidemiology

Vol. 192, No. 3
https://doi.org/10.1093/aje/kwac067
Advance Access publication:
April 6, 2022

Practice of Epidemiology

Illustration of 2 Fusion Designs and Estimators

Stephen R. Cole*, Jessie K. Edwards, Alexander Breskin, Samuel Rosin, Paul N. Zivich, Bonnie E. Shook-Sa, and Michael G. Hudgens

* Correspondence to Dr. Stephen Cole, Department of Epidemiology, Gillings School of Global Public Health, University of North Carolina, Campus Box 7435, Chapel Hill, NC 27599-7435 (e-mail: cole@unc.edu).

Initially submitted May 10, 2021; accepted for publication March 31, 2022.

"Fusion" study designs combine data from different sources to answer questions that could not be answered (as well) by subsets of the data. Studies that augment main study data with validation data, as in measurement-error correction studies or generalizability studies, are examples of fusion designs. Fusion estimators, here solutions to stacked estimating functions, produce consistent answers to identified research questions using data from fusion designs. In this paper, we describe a pair of examples of fusion designs and estimators, one where we generalize a proportion to a target population and one where we correct measurement error in a proportion. For each case, we present an example motivated by human immunodeficiency virus research and summarize results from simulation studies. Simulations demonstrate that the fusion estimators provide approximately unbiased results with appropriate 95% confidence interval coverage. Fusion estimators can be used to appropriately combine data in answering important questions that benefit from multiple sources of information.

accuracy; bias; generalizability; measurement error; random error; study design

Abbreviations: AIDS, acquired immunodeficiency syndrome; CI, confidence interval; HIV, human immunodeficiency virus; SE, standard error.

Editor's note: *An invited commentary on this article will appear in a future issue.*

A "fusion" study design combines data from multiple sources to answer a question that could not be answered (as well) by data from subsets of these sources (1). For a question to be answered, the parameter which addresses the question must be (partially or point) identified given the data observed under the study design. Fusion designs may include data on participants from different studies or on participants from different stages of a single (nested) study. Examples of the former type include measurement-error correction studies which use external validation data, generalizability studies which use auxiliary information on the target population (2), and studies of bridged treatment effects which rely on a set of treatment comparison studies (3). Examples of the latter type include measurement-error correction studies which use internal validation data, generalizability studies where trials are nested in cohorts (4), and classical 2-stage studies which collect costly covariate information on a subset of participants (5, 6). Here we focus on examples of the former type with external auxiliary data.

A fusion estimator produces an answer to a question using data from a fusion study. Here, we consider fusion estimators that are solutions to stacked estimating functions (7, 8). It is desirable that such estimators converge in probability to the parameter value as the sample size increases (i.e., consistent), converge in distribution to a Gaussian random variable (i.e., asymptotically normal), and are precise enough to be useful in practice, even if not optimally efficient (8). In the analysis of fusion designs, it is important to appropriately propagate the uncertainty from the various data sources when estimating the parameter(s) of interest. Some widely used approaches to quantifying uncertainty associated with parameter estimates from fusion designs require intense computation. For example, one might use the bootstrap to estimate the random error through repeated sampling from the set of data sources (9) or employ a Bayesian approach using Markov chain Monte Carlo sampling (10).

Our motivation here is 2-fold. First, framing studies with multiple data sources as fusion designs helps to connect apparently disparate approaches and clarify conditions sufficient to identify parameters of interest in diverse settings. Second, using well-established estimating function-based approaches to obtain parameter estimates for fusion designs is highly flexible and more computationally efficient than some competing approaches, while still appropriately propagating random error. The present work introduces epidemiologists to some key ideas in fusion designs and a convenient approach to estimation. Below we describe a pair of minimal nontrivial examples of fusion designs and estimators. First, an estimate of the proportion is transported using auxiliary data from a target population. Second, an estimate of the proportion based on misclassified data is corrected using auxiliary validation data. For each example, the methods are illustrated with data from human immunodeficiency virus (HIV) research and simulation studies are conducted to explore the finite sample properties of the estimators illustrated.

METHODS

Identification conditions

Here we show how to correct for biases in epidemiologic studies that leverage multiple sources of data. Below, in case 1, we generalize a proportion to a target population, and in case 2 we correct measurement error in a proportion. In case 1, we will have a sample, indicated as $R = 0$, which includes only the variable W (playing the role of the population-characterizing covariate). We will also have a sample, indicated as $R = 1$, which includes variables W and Y. The stacked data X_i can be written as $\{W_i, Y_i R_i, R_i\}$ for $i = 1, \ldots, N$. In case 2, we will again have a sample, indicated as $R = 0$, which only includes variable W (now playing the role of mismeasured Y). We will also have a sample, indicated as $R = 1$, which includes variables W and Y. In case 2, the $R = 1$ sample can be split into subsamples defined by $Y = 1$ and $Y = 0$, as done below. The stacked data X_i can again be written as $\{W_i, Y_i R_i, R_i\}$. The target parameter for both case 1 and case 2 is $P(Y = y|R = 0)$.

In case 1, the 2 identification conditions are $P(Y = y|W = w, R = 0) = P(Y = y|W = w, R = 1)$ and $P(W = w|R = 1) > 0$ for all w, where $P(W = w|R = 0) > 0$, assuming for ease of notation that W is discrete (an analogous condition holds when W is continuous). The first identification assumption is a conditional exchangeability statement, such that the outcome Y is missing at random given W. We must select W on the basis of expert knowledge to control selection bias. The second identification assumption is a related positivity statement.

For case 1, we reexpress the parameter of interest by

$$P(Y = y|R = 0) = \sum_w P(Y = y|W = w, R = 0)\, P(W = w|R = 0)\,. \quad (1)$$

The first term on the right side of equation 1 is generally not identifiable without assumptions, because Y is not observed when $R = 0$. However, under the above identification assumptions, this term can be replaced by $P(Y = y|W = w, R = 1)$, such that

$$P(Y = y|R = 0) = \sum_w P(Y = y|W = w, R = 1)\, P(W = w|R = 0)\,, \quad (2)$$

where the right side of equation 2 is now comprised solely of observed quantities. Specifically, $P(Y = y|W = w, R = 1)$ is identified from the $R = 1$ sample, and $P(W = w|R = 0)$ is identified from the $R = 0$ sample because in the $R = 0$ sample we observe W and these data are assumed to be a random sample from the target population.

In case 2, we reexpress the parameter of interest as

$$P(Y = 1|R = 0) = \frac{P(W = 1|R = 0) + P(W = 0|Y = 0, R = 0) - 1}{P(W = 1|Y = 1, R = 0) + P(W = 0|Y = 0, R = 0) - 1}. \quad (3)$$

The $R = 0$ sample provides a random sample of data on W from the target population to identify $P(W = 1|R = 0)$, and the $R = 1$ sample identifies the other probabilities on the right side of equation 3 under the exchangeability condition $P(W = w|Y = y, R = 1) = P(W = w|Y = y, R = 0)$. These same identification conditions may similarly be adapted to other cases of fusion designs.

M-estimators

Here we provide a brief review of M-estimation. For more detail, see Godambe (7) and Stefanski and Boos (8). Many epidemiologic analyses entail estimating a vector parameter $\theta = (\theta_1, \ldots, \theta_p)$ using data from N independent individuals $X_1, \ldots, X_N$. Often the estimator $\hat{\theta}$ of θ can be expressed as the solution to an estimating equation $\sum_{i=1}^{N} g(X_i; \theta) = 0$, where g is a (column) vector of real-valued functions of the data X_i and the parameters θ. For example, if θ were the expected value of X, and $\hat{\theta}$ was the sample mean $N^{-1}\sum_{i=1}^{N} X_i$, then $g(X_i; \theta)$ would equal $X_i - \theta$. Maximum likelihood estimators are M-estimators, where the score equations (i.e., partial derivatives of the log likelihood) are used for $g(X_i, \theta)$ (the "M" in "M-estimator" reflects the fact that many estimators maximize some objective function, or equivalently, equal a root of the derivative of that function). If the estimating function $g(X; \theta)$ is unbiased (has expectation 0 at the true value of θ) and suitable regularity conditions hold (see Boos and Stefanski (11), chapter 7), then as the sample size N increases, the M-estimator $\hat{\theta}$ will be consistent and asymptotically normal (8). The asymptotic variance of $\hat{\theta}$ can be estimated by the empirical sandwich estimator $\Sigma_{\hat{\theta}} = B_N(\hat{\theta})^{-1} M_N(\hat{\theta}) B_N(\hat{\theta})^{-T}$. The bread of the sandwich

estimator is $B_N(\hat{\theta}) = N^{-1}\sum_{i=1}^{N} g'(X_i;\hat{\theta})$, where $g'(X_i;\theta) = -\left[\frac{\partial g(X_i;\theta)}{\partial \theta}\right]$ is the matrix of derivatives of the estimating function with respect to the parameters. These derivatives may be derived analytically or approximated numerically. The meat is $M_N(\hat{\theta}) = N^{-1}\sum_{i=1}^{N} g(X_i;\hat{\theta})g(X_i;\hat{\theta})^T$—that is, the sample average of the outer product of the estimating function. The asymptotic normality of the estimator $\hat{\theta}$ justifies use of Wald-type confidence intervals (CIs) in large samples. Here, the SAS procedure IML was used for M-estimation (SAS Institute Inc., Cary, North Carolina). Software code for example 2 is provided in Web Appendix 1 (available at https://doi.org/10.1093/aje/kwac067). Alternatively, the package "geex" (12) can be used to compute M-estimators and empirical sandwich variance estimators in R (R Foundation for Statistical Computing, Vienna, Austria).

CASE 1: TRANSPORTING THE PROPORTION

Approach

Say a study sample of n units is available, with binary outcome Y and measured covariates W, with records denoted by $R = 1$. The goal is to estimate the prevalence of Y in a target population using data from this study. However, the study sample is not a random sample from the target population; rather, it is a biased sample from the target population (or can be thought of as a random sample from some other population). We also have a random sample of m additional units from the target population, with measured covariates W (but the outcome Y is not measured in this auxiliary data), with records denoted by $R = 0$. The n and m units are independent (both within and across samples) but are not identically distributed. Here $N = n + m$.

Using only data from the $R = 1$ sample will generally result in a biased estimate of the prevalence of Y in the target population. The $R = 0$ random sample from the target population provides a way to correct for this bias. For intuition, think of the $R = 1$ sample as a W-stratified random sample from the target population, where the strata-specific sampling probabilities are unknown. The $R = 0$ random sample from the target population provides a way to recover the unknown sampling probabilities. Specifically, we estimate a "sampling" score as the probability of a unit appearing in the main study, versus the auxiliary study, given that the unit appears in one or the other, or $P(R = 1|W)$, where $R = 1$ if the unit is in the main study and $R = 0$ if the unit is in the auxiliary study. Then we use the predicted probabilities from the sampling score model to estimate the sampling odds weights, $P(R = 0|W)/P(R = 1|W)$, and weight individuals in the $R = 1$ sample so that, asymptotically, the weighted sample has the same W distribution as the target population.

Momentarily assume the sampling odds weights are known. Let $\theta = E(Y|R = 0)$ denote the mean of Y in the target population, $X = (YR, R, W)$ the observable data, and $\pi = \pi(W) = P(R = 0|W)/P(R = 1|W)$. Note that the estimating function for θ, $g(X;\theta) = R(Y - \theta)\pi$, has mean 0 (Web Appendix 2). Therefore, a consistent and asymptotically normal estimator of the mean of Y can be obtained by solving $\sum_{i=1}^{N} g(X_i;\theta) = 0$, which has the closed-form solution $\hat{\theta} = \sum_{i=1}^{N} R_i Y_i \pi_i / \sum_{i=1}^{N} R_i \pi_i$ where $\pi_i = \pi(W_i)$.

In practice, the odds weights are typically not known and instead are estimated. To estimate the weights, assume the logistic model $P(R = 1|W;\beta) = [\exp(-\beta W^*) + 1]^{-1}$, where $W^* = (1, W)^T$ and the row vector parameter β has the same dimension as W^{*T}. Let the mean of Y in the target population now be denoted as μ and let $\theta = (\mu, \beta)$. A consistent estimator $\hat{\theta}$ can be obtained by finding the value of θ that solves $\sum_{i=1}^{N} g_\theta(X_i;\theta) = 0$, where

$$g_\theta(X;\theta) = \begin{pmatrix} g_{\mu,\beta}(Y, R, W;\mu,\beta) \\ g_\beta(R, W;\beta) \end{pmatrix} = \begin{pmatrix} R[Y - \mu]\pi \\ [R - P(R = 1|W;\beta)]W^* \end{pmatrix};$$

here, $\pi = \pi(W;\beta) = P(R = 0|W;\beta)/P(R = 1|W;\beta)$. In general, the estimating equation solution $\hat{\theta}$ may be found numerically through a root-finding algorithm, such as Newton's method (13). The variance-covariance matrix of $\hat{\theta}$ can be estimated by $\Sigma_{\hat{\theta}}/N$, where the empirical sandwich estimator $\Sigma_{\hat{\theta}}$ is computed on the basis of the vector estimating function $g_\theta(X;\theta)$. See Web Appendix 3 for additional details on the sandwich estimator for this case. For this problem, $\hat{\theta}$ can also be obtained by finding the maximum likelihood estimates of the logistic sampling model parameters, say $\hat{\beta}$, and estimating μ via the closed-form estimator in the preceding paragraph, with $\pi_i = \pi(W_i;\hat{\beta})$.

Example

We wish to learn the 1-year risk of acquired immunodeficiency syndrome (AIDS) or death among HIV-positive adults placed on an HIV treatment plan in the United States in the late 1990s. In this main study, we have a sample of $n = 579$ HIV-positive adults randomized to this treatment plan, as described in the 2-drug arm of Hammer et al.'s (14) Table 2. In this study, 63 (11%) of the 579 patients developed AIDS or died during the 1-year follow-up period. Among these participants, 94 (16%) of the 579 were female, and 388 (67%) were aged ≥35 years.

Auxiliary data come from a Centers for Disease Control and Prevention census of the target population conducted in 2006, including the characteristics sex and age group (see Cole and Stuart (2)). Of the 121,617 diagnosed HIV cases, 20,430 (17%) were female and 63,689 (52%) were aged ≥35 years.

To simplify presentation, we treat the 31 (5%) patients in the main sample of 579 who were lost to follow-up as if they had no events. To amplify the difference between the sample estimate and the transported estimate for demonstration purposes, we reclassify 14 of the people with outcomes in the main sample who were aged 18–34 years as outcomes

Table 1. Estimated Bias, Standard Errors, and 95% Confidence Interval Coverage for 5,000 Simulations Transporting a Proportion to a Given Target Population

Biased Main-Study Sample Size, *n*	Auxiliary Random Sample Size, *m*	True Proportion	Bias[a]	Average Sandwich SE	Empirical SE	95% CI Coverage
200	200	0.1	−0.002	0.020	0.021	93.3
	200	0.5	0.007	0.038	0.039	94.0
	500	0.1	0.000	0.020	0.020	93.9
	500	0.5	0.011	0.037	0.037	93.8
500	200	0.1	−0.003	0.014	0.014	93.5
	200	0.5	0.005	0.027	0.026	94.2
	500	0.1	−0.002	0.013	0.013	94.0
	500	0.5	0.009	0.024	0.024	93.2

Abbreviations: CI, confidence interval; SE, standard error.
[a] Monte Carlo simulation SE $\leq$ 0.001 for each row.

among those aged $\geq$35 years. To illustrate features of the proposed approach, we take simple random samples without replacement of sizes $m = 200$, $m = 500$, and $m = 5{,}000$ from the auxiliary Centers for Disease Control and Prevention data. The logistic model fitted to estimate the sampling score included sex, age $\geq$35 years, and their product.

Example results

The observed outcome proportion in the main sample was 10.9% (63/579). The outcome was more common among patients aged $\geq$35 years (i.e., 61/402) than among those aged 18–34 years (i.e., 2/177). With an auxiliary sample size of 200, the estimated transported proportion who developed AIDS or died was 8.0% (95% CI: 5.8, 10.1). Given the structure of this example, we would expect the transported estimate to be smaller than the sample estimate. In the main sample of 579, the estimated odds weights had a mean of 0.35, with a minimum of 0.23 and a maximum of 0.61. The sandwich standard error (SE) for the proportion developing AIDS or death was 0.0109, the sandwich SE that ignores the variability in the weights was 0.0102, and the nonparametric bootstrap SE was 0.0117 (the nonparametric bootstrap SE was calculated by the standard deviation of 500 estimates, each estimate based on simple random samples of sizes n and m with replacement from the observed data sets).

With an auxiliary sample size of 500, roughly the same as the size of the main study sample, the estimated proportion who developed AIDS or died was 7.8% (95% CI: 5.8, 9.8). In the main sample of 579, the odds weights had a mean of 0.86, with a minimum of 0.57 and a maximum of 1.58. Here, the sandwich SE for the proportion who developed AIDS or died was 0.0101, the sandwich SE that ignored the variability in the weights was also 0.0101, and the nonparametric bootstrap SE was 0.0105.

With a large auxiliary sample size of 5,000, the estimated proportion who developed AIDS or died was 8.5% (95% CI: 6.5, 10.5). In the main sample of 579, the odds weights had a mean of 8.6, with a minimum of 6.3 and a maximum of 14.7. Here, the sandwich SE for the proportion who developed AIDS or died was 0.0102, the sandwich SE that ignored the variability in the weight was 0.0107, and the nonparametric bootstrap SE was 0.0107. In summary, regardless of the size of the auxiliary data, accounting for biased sampling yielded a clinically meaningful 2%–3% difference in the estimated proportion who developed AIDS or died.

Simulations

We explored 8 scenarios, one of which was chosen to roughly mimic the example. We generated 5,000 biased main study samples, each of size 200 or 500, with a true proportion of Y of 10% or 50%, and a standard normal auxiliary covariate W. Specifically, Y was a Bernoulli random variable with marginal expectation of 0.1 or 0.5 and an odds ratio of e^1 for a unit difference in W. The indicator of sample selection R was also a Bernoulli random variable with marginal expectation of 0.5 and an odds ratio of e^1 for a unit difference in W. We also generated 5,000 simple random auxiliary samples of size 200 or 500, with only W measured. For each of the scenarios explored, we present the bias (i.e., the average of the difference between the estimates and the true data-generating value), average sandwich SE, empirical SE (i.e., standard deviation of simulated point estimates), and 95% CI coverage (i.e., proportion of 95% CIs that contain the true value).

Simulation results

Table 1 shows the bias, average sandwich SE, empirical SE, and 95% CI coverage. The fusion estimator was approximately unbiased, with an average (over 8 scenarios) of the absolute value of the bias of 0.005. The average sandwich SE approximated the empirical SE well across the scenarios explored. The average (over 8 scenarios) 95% CI coverage was 93.7%.

CASE 2: ESTIMATING A MISCLASSIFIED PROPORTION

Approach

Say we want to estimate the proportion exposed in a population at a point in time, specifically, $P(Y = 1|R = 0) = \alpha$, where Y is the binary exposure indicator variable. Note that symbols are recycled, such that the notation from case 1 does not carry across here to case 2. We have a random sample of n_0 units from the population of interest, with binary W being a mismeasured version of Y. We have a random sample of n_1 units, with W measured, from the stratum of the population where $Y = 1$, and we have a random sample of n_2 units, with W measured, from the stratum of the population where $Y = 0$.

It is straightforward to show (15, 16) that $P(Y = 1|R = 0) = [P(W = 1|R = 0) + \delta - 1]/(\gamma + \delta - 1)$, where $\gamma = P(W = 1|Y = 1, R = 0)$ is the sensitivity and $\delta = P(W = 0|Y = 0, R = 0)$ is the specificity of the measurement instrument. Let $R_i = 0$ if unit i is one of the n_0 units from the main sample, $R_i = 1$ if unit i is one of the n_1 units from the auxiliary sample where $Y = 1$, and $R_i = 2$ if unit i is one of the n_2 units from the auxiliary sample where $Y = 0$. Then $X = (W, R)$, $N = n_0 + n_1 + n_2$, and the stacked estimating function is

$$g_\theta(X;\theta) = \begin{pmatrix} g_\beta(X;\beta) \\ g_\gamma(X;\gamma) \\ g_\delta(X;\delta) \\ g_\alpha(X;\alpha,\beta,\gamma,\delta) \end{pmatrix} = \begin{pmatrix} I(R=0)\,(W-\beta) \\ I(R=1)\,(W-\gamma) \\ I(R=2)\,[(1-W)-\delta] \\ \alpha(\gamma+\delta-1)-(\beta+\delta-1) \end{pmatrix},$$

where $\theta = (\alpha, \beta, \gamma, \delta)$. Note that the last row of $g_\theta(X;\theta)$ is a function only of the other parameters (see example 2 in Stefanski and Boos (8)).

Example

We wish to learn the point prevalence of HIV treatment for HIV-positive adults in the United States. We have a sample of $n_1 = 950$ HIV-positive adults enrolled in HIV interval cohort studies, as described in Cole et al.'s (17) Table 1. Data from independent and exchangeable patients are available for sensitivity ($n_2 = 242$) and specificity ($n_3 = 89$) of the self-report of HIV treatment as compared with extensive medical and pharmacy records review as a gold standard (see Appendix of Cole et al. (17)). To illustrate some features of the proposed approach, we consider a separate example that assumes the auxiliary sample data set has the size 20,000 (= 10,000 + 10,000), rather than the observed 331 (= 242 + 89).

Example results

If 680 of the 950 participants report HIV treatment, the observed prevalence is 72% (95% CI: 69, 74). Given the auxiliary data on 331 individuals, the estimated sensitivity and specificity are 84% (95% CI: 80, 89) and 80% (95% CI: 71, 88), respectively. The fusion estimator of the prevalence, accounting for imperfectly known sensitivity and specificity, is 80% (95% CI: 72, 88). The fusion estimator is notably larger than the naive observed prevalence of 72% because of the modest sensitivity and specificity. The sandwich SE estimate for the fusion estimator is 0.040, while the estimated SE for the observed data estimator is 0.015, illustrating the uncertainty added by the fusion estimator as a cost of correcting the point estimate with data from relatively small validation samples. To compare, we also estimated the SE of the corrected proportion using the standard deviation of 500 nonparametric bootstrap random samples (size n_1, n_2, n_3 with replacement). The bootstrap SE estimate was 0.039.

If the true prevalence were indeed 80% (i.e., the fusion estimate was unbiased), the root mean squared error for the fusion estimator would be 0.040, while the root mean squared error for the observed data estimator would be 0.081. Under this assumption, the sizable measurement bias reduction outweighs the added uncertainty, at least in terms of squared error.

If the validation sample were of size 20,000 rather than 331, then both the sensitivity and the specificity would be subject to much less random error, and the estimated sandwich SE for the fusion estimator would be reduced from 0.040 to 0.023 (and the bootstrap SE estimate would be 0.022).

Simulations

Data were simulated under 18 scenarios, 1 of which was chosen to approximately mimic the example. For each scenario, 5,000 samples were generated, each of size 1,000. The scenarios were defined by a true point prevalence of 50% or 80%; an observed point prevalence misclassified with sensitivity and specificity of 85% and 80%, 85% and 75%, or 75% and 75%, respectively; and independent auxiliary samples for sensitivity and specificity of sizes 200 and 200, 200 and 100, or 100 and 100, respectively.

Simulation results

Table 2 shows the bias, average sandwich SE, empirical SE, and coverage of the 95% CI. The upper half of Table 2 details results for situations where the true prevalence is 50%. The bottom half of Table 2 details results for situations where the true prevalence is 80%. The method is approximately unbiased for the prevalence, and the sandwich SE estimator is also approximately unbiased. The 95% CI provides approximately nominal coverage, with an average coverage of 95.9% for the 18 scenarios in Table 2.

DISCUSSION

Here we have demonstrated 2 simple examples of fusion designs and estimators. In both cases, the examples illustrated the benefit of the fusion estimator, and the simulations showed favorable operating characteristics.

Table 2. Estimated Bias, Standard Errors, and 95% Confidence Interval Coverage for 5,000 Simulations Correcting a Misclassified Proportion

Sensitivity	Specificity	Validation Sizes[a]	Bias[b]	Average Sandwich SE	Empirical SE	95% CI Coverage
			Prevalence = 0.5			
0.85	0.80	200, 200	0.000	0.039	0.038	95.5
	0.80	200, 100	−0.001	0.045	0.045	95.6
	0.80	100, 100	0.000	0.050	0.048	96.3
	0.75	200, 200	−0.001	0.043	0.043	95.8
	0.75	200, 100	−0.002	0.051	0.050	95.7
	0.75	100, 100	−0.001	0.055	0.054	96.3
0.75	0.75	200, 200	0.000	0.056	0.055	95.8
	0.75	200, 100	−0.002	0.064	0.064	96.4
	0.75	100, 100	0.000	0.072	0.070	96.7
			Prevalence = 0.8			
0.85	0.80	200, 200	0.001	0.040	0.039	95.4
	0.80	200, 100	0.000	0.041	0.040	95.8
	0.80	100, 100	0.002	0.053	0.051	96.1
	0.75	200, 200	0.000	0.044	0.043	95.6
	0.75	200, 100	0.000	0.045	0.044	96.0
	0.75	100, 100	0.002	0.058	0.056	96.2
0.75	0.75	200, 200	0.002	0.060	0.061	95.3
	0.75	200, 100	0.001	0.061	0.061	95.7
	0.75	100, 100	0.004	0.080	0.080	95.8

Abbreviations: CI, confidence interval; SE, standard error.
[a] Validation sizes are for sensitivity and specificity samples; the main study sample size was 1,000.
[b] Monte Carlo simulation SE $\leq$ 0.001 for each row.

Fusion designs are common. There is a long history of combining empirical studies, in epidemiology, statistics, and elsewhere (e.g., see Pearson (18)). Recent work on generalizability can be cast as fusion designs (2, 4, 19, 20), and early examples of fusion designs include 2-stage studies (5, 6, 21). For example, in a classical 2-stage study, covariate information is available only for a subset of study participants. Likewise, in some measurement-error correction and generalizability studies, the auxiliary data are obtained on a subset of participants in a single study rather than from multiple studies; but in such cases, the stages of the study can be thought of as distinct data sources to facilitate the use of fusion designs. Indeed, in such cases where the auxiliary data are obtained on a *random* subsample, the identification conditions given above are guaranteed by design, rather than by assumption. The fusion estimators described here can be adapted to such settings (22). There are also connections between meta-analysis and fusion designs. Traditionally, meta-analysis has entailed combining trial-specific effect estimates; however, the resulting pooled estimate does not necessarily have a causal interpretation. On the other hand, valid causal inferences can be drawn by pooling, or fusing, data from individual trials (23). Although not illustrated in this paper, fusion estimators can be semiparametric, when 1 or more components of the parameter θ are infinite dimensional and the balance are finite dimensional (24, 25). Likewise, fusion estimators can be (semi-) Bayes estimators, by leveraging existing information for (a subset of) the parameters (26–28).

There are limitations to fusion designs and estimators. First, some identification conditions must be met in any study design for an estimator to provide valid results. Indeed, there are additional identification conditions for fusion study designs regarding the exchangeability between data sources. For example, see the discussion of identification in recent work on generalizability (29) and bridged treatment comparisons (3). While fusion designs can identify parameters not identified in conventional study designs, further research is needed regarding the extent to which fusion design identification conditions have testable implications (23, 30). Second, the estimators illustrated here have a large sample justification and are subject to finite sample bias. In small samples, penalization may improve operating characteristics (31, 32). Third, here, in both cases, we accounted for estimation of the nuisance parameters when estimating the variance of the parameters of interest. One need not always do so. Newey and McFadden showed that we generally must account for estimation of the nuisance parameters if and only if consistency of the nuisance parameter estimator affects consistency of the target parameter estimator (see Newey

and McFadden (33), section 6). Fourth, the inverse odds weighted estimator described for case 1 is not semiparametric efficient. A more efficient estimator can be obtained by augmenting this estimator with information from an outcome model (24, 34). Fifth, when using finite dimension parametric models, as in case 1, when such a model is misspecified the resulting estimate may be interpreted as estimating the least false parameter. In such cases, bias will result when the least false parameter does not coincide with the population parameter. Sixth and last, there is much still to understand regarding the use of fusion designs in epidemiology. Comparisons of the accuracy (e.g., mean squared error) of parameter estimators from standard and fusion study designs, across a broad set of realistic scenarios, would be helpful.

In conclusion, fusion study designs can help investigators appropriately combine data from multiple sources of information. A unification of this variety of designs may provide hidden insights and allow epidemiologists to better realize the power of combining data to answer important questions. Moreover, such a unification would enhance communication and may help create a common language for individuals from different disciplines to make better use of seemingly discipline-specific study designs.

ACKNOWLEDGMENTS

Author affiliations: Department of Epidemiology, Gillings School of Global Public Health, University of North Carolina, Chapel Hill, North Carolina, United States (Stephen R. Cole, Jessie K. Edwards, Paul N. Zivich); NoviSci LLC, Durham North Carolina, United States (Alexander Breskin); and Department of Biostatistics, Gillings School of Global Public Health, University of North Carolina, Chapel Hill, North Carolina, United States (Samuel Rosin, Bonnie E. Shook-Sa, Michael G. Hudgens).

This work was supported in part by National Institutes of Health grants R01AI157758 (S.R.C., J.K.E., M.G.H.), P30AI50410 (S.R.C., M.G.H.), K01AI125087 (J.K.E.), and T32AI007001 (P.N.Z.) and National Science Foundation grant NSF DGE-1650116 (S.R.).

Data sets are available from the corresponding author.

The views expressed in this article are those of the authors and do not reflect those of the National Institutes of Health.

Conflict of interest: none declared.

REFERENCES

1. Bareinboim E, Pearl J. Causal inference and the data-fusion problem. *Proc Natl Acad Sci U S A.* 2016;113(27):7345–7352.
2. Cole SR, Stuart EA. Generalizing evidence from randomized clinical trials to target populations: the ACTG 320 Trial. *Am J Epidemiol.* 2010;172(1):107–115.
3. Breskin A, Cole SR, Edwards JK, et al. Fusion designs and estimators for treatment effects. *Stat Med.* 2021;40(13):3124–3137.
4. Dahabreh IJ, Robertson SE, Steingrimsson JA, et al. Extending inferences from a randomized trial to a new target population. *Stat Med.* 2020;39(14):1999–2014.
5. Walker AM. Anamorphic analysis: sampling and estimation for covariate effects when both exposure and disease are known. *Biometrics.* 1982;38(4):1025–1032.
6. White JE. A two stage design for the study of the relationship between a rare exposure and a rare disease. *Am J Epidemiol.* 1982;115(1):119–128.
7. Godambe VP. *Estimating Functions*. New York, NY: Clarendon Press; 1991.
8. Stefanski LA, Boos DD. The calculus of M-estimation. *Am Stat.* 2002;56(1):29–38.
9. Greenland S. Interval estimation by simulation as an alternative to and extension of confidence intervals. *Int J Epidemiol.* 2004;33(6):1389–1397.
10. Gustafson P. *Measurement Error and Misclassification in Statistics and Epidemiology: Impacts and Bayesian Adjustments*. New York, NY: Chapman & Hall/CRC Press; 2003.
11. Boos DD, Stefanski LA. *Essential Statistical Inference*. New York, NY: Springer Publishing Company; 2013.
12. Saul BC, Hudgens MG. The calculus of M-estimation in R with geex. *J Stat Softw.* 2020;92(2):1–15.
13. Hamming RW. *Numerical Methods for Scientists and Engineers*. 2nd ed. Mineola, NY: Dover Publications; 1986:68.
14. Hammer SM, Squires KE, Hughes MD, et al. A controlled trial of two nucleoside analogues plus indinavir in persons with human immunodeficiency virus infection and CD4 cell counts of 200 per cubic millimeter or less. AIDS Clinical Trials Group 320 Study Team. *N Engl J Med.* 1997;337(11):725–733.
15. Rogan WJ, Gladen B. Estimating prevalence from the results of a screening test. *Am J Epidemiol.* 1978;107(1):71–76.
16. Greenland S. Basic methods for sensitivity analysis of biases. *Int J Epidemiol.* 1996;25(6):1107–1116.
17. Cole SR, Jacobson LP, Tien PC, et al. Using marginal structural measurement-error models to estimate the long-term effect of antiretroviral therapy on incident AIDS or death. *Am J Epidemiol.* 2010;171(1):113–122.
18. Pearson K. Report on certain enteric fever inoculation statistics. *Br Med J.* 1904;3(2288):1243–1246.
19. Buchanan AL, Hudgens MG, Cole SR, et al. Generalizing evidence from randomized trials using inverse probability of sampling weights. *J R Stat Soc Ser A Stat Soc.* 2018;181(4):1193–1209.
20. Dahabreh IJ, Haneuse SJA, Robins JM, et al. Study designs for extending causal inferences from a randomized trial to a target population. *Am J Epidemiol.* 2021;190(8):1632–1642.
21. Neyman J. On two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. *J R Stat Soc.* 1934;97:558–625.
22. Flanders WD, Greenland S. Analytic methods for two-stage case-control studies and other stratified designs. *Stat Med.* 1991;10(5):739–747.
23. Dahabreh IJ, Petito LC, Robertson SE, et al. Toward causally interpretable meta-analysis: transporting inferences from multiple randomized trials to a new target population. *Epidemiology.* 2020;31(3):334–344.
24. Robins JM, Rotnitzky A, Zhao LP. Estimation of regression coefficients when some regressors are not always observed. *J Am Stat Assoc.* 1994;89(427):846–846.

25. Tsiatis AA. *Semiparametric Theory and Missing Data*. New York, NY: Springer Publishing Company; 2006.
26. Greenland S. A semi-Bayes approach to the analysis of correlated multiple associations, with an application to an occupational cancer-mortality study. *Stat Med.* 1992;11(2): 219–230.
27. Good IJ. The Bayes/non-Bayes compromise: a brief review. *J Am Stat Assoc.* 1992;87(419):597–606.
28. Godambe VP. Estimating functions: a synthesis of least squares and maximum likelihood methods. *IMS Lecture Notes Monogr Ser.* 1997;32:5–16.
29. Lesko CR, Buchanan AL, Westreich D, et al. Generalizing study results: a potential outcomes perspective. *Epidemiology.* 2017;28(4):553–561.
30. Stuart EA, Cole SR, Bradshaw CP, et al. The use of propensity scores to assess the generalizability of results from randomized trials. *J R Stat Soc Ser A Stat Soc.* 2011;174(2): 369–386.
31. Cole SR, Chu H, Greenland S. Maximum likelihood, profile likelihood, and penalized likelihood: a primer. *Am J Epidemiol.* 2014;179(2):252–260.
32. Cole SR, Edwards JK, Westreich D, et al. Estimating multiple time-fixed treatment effects using a semi-Bayes semiparametric marginal structural Cox proportional hazards regression model. *Biom J.* 2018;60(1):100–114.
33. Newey WK, McFadden D. Large sample estimation and hypothesis testing. In: Engle RF, McFadden D, eds. *Handbook of Econometrics.* (Vol. 4). New York, NY: Elsevier B.V.; 1994:2111–2245.
34. Cole SR, Edwards JK, Breskin A, et al. Comparing parametric, nonparametric, and semiparametric estimators: the Weibull trials. *Am J Epidemiol.* 2021;190(8):1643–1651.